

High-resolution coarse-grained modeling using oriented coarse-grained sites

Thomas K. Haxton

Molecular Foundry, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA

We introduce a method to bring nearly atomistic resolution to coarse-grained models, and we apply the method to proteins. Using a small number of coarse-grained sites (about one per eight atoms) but assigning an *independent three-dimensional orientation* to each site, we preferentially integrate out stiff degrees of freedom (bond lengths and angles, as well as dihedral angles in rings) that are accurately approximated by their average values, while retaining soft degrees of freedom (unconstrained dihedral angles) mostly responsible for conformational variability. We demonstrate that our scheme retains nearly atomistic resolution by mapping all experimental protein configurations in the Protein Data Bank onto coarse-grained configurations, then analytically backmapping those configurations back to all-atom configurations. This roundtrip mapping throws away all information associated with the eliminated (stiff) degrees of freedom except for their average values, which we use to construct optimal backmapping functions. Despite the 4:1 reduction in the number of degrees of freedom, we find that heavy atoms move only 0.051 Å on average during the roundtrip mapping, while hydrogens move 0.179 Å on average, an unprecedented combination of efficiency and accuracy among coarse-grained protein models. We discuss the advantages of such a high-resolution model for parameterizing effective interactions and accurately calculating observables through direct or multiscale simulations.

The development of accurate classical force-fields and special-purpose supercomputers have allowed all-atom molecular simulations to provide a rich microscopic view of many complex processes in molecular biology. For example, all-atom molecular dynamics simulations are now able to explore in atomic detail the folding kinetics of fast-folding protein domains [1, 2], engineered protein dimers [3], and small, naturally-occurring proteins [4], using simulations spanning milliseconds for proteins containing up to 100 residues.

While hardware and software advances will continue to advance the dynamic range of all-atom simulations, accessible time and length scales will always be limited by the short time steps (typically 10^{-15} s) required to resolve atomic vibrational motion [5] and the computational expense of propagating all the atomic degrees of freedom. Thus, many important biological processes remain significantly beyond the frontier of all-atom simulations. For example, understanding complex and competing assembly pathways of soluble and insoluble aggregates is crucial for developing therapies for Alzheimer's disease, but these pathways can each involve ten or more 40-residue peptides and span hours or days [6–8].

Coarse-grained modeling presents an appealing way to extend the time and length scales accessible to molecular simulations [9–20]. By reducing the number of degrees of freedom by a factor of N , coarse-grained simulations decrease the number of short-range force calculations and increase accessible time scales both by factors of N^2 . Averaging out the stiffest degrees of freedom allows longer time steps in molecular dynamics simulations (or increased trial step sizes in Monte Carlo simulations), further increasing the accessible dynamic range. Coarse-grained models can be used by themselves or as components in multiscale simulation schemes [21–33] that use coarse-grained simulations to accelerate sampling or dynamics of related all-atom simulations.

Information is lost when integrating out degrees of freedom, so the increased efficiency of coarse-grained modeling is always paid for with a decrease in accuracy [34–38]. Because complex biological systems are sensitive to decreases in accuracy, coarse-grained modeling has in many cases failed in a qualitative way to extend the reach of molecular simulation. For example, unbiased coarse-grained models have not yet successfully folded any proteins that cannot be folded by all-atom simulations. Thus, methods to improve the accuracy of coarse-grained models are urgently needed.

The accuracy of coarse-grained models can be improved by optimizing one or both of the two aspects of coarse-grained modeling: (1) the mapping between all-atom and coarse-grained degrees of freedom and (2) the effective interactions among the coarse-grained degrees of freedom. Many authors have focused on the second aspect, developing efficient methods to parameterize interactions so that differences between coarse-grained simulations and reference all-atom simulations are minimized. Examples include iterative Boltzmann inversion that matches pair distribution functions [39], the multiscale coarse-graining method that matches forces [40–43], and the relative entropy method that minimizes the relative entropy between coarse-grained and all-atom ensembles [44–47].

Some authors have also begun exploring different ways to map atomic positions onto coarse-grained sites, which in turn dictates the types of effective interactions available. Most coarse-grained models are point-site models, mapping groups of atoms onto structureless coarse-grained sites. In many point-site models, the effective interactions are analogous to all-atom forcefields: interactions between covalently connected sites are decomposed into bond stretching, bending, and twisting terms, and non-bonded interactions are spherically symmetric. Although point-site models discard all information about

the structure of each site’s associated atoms, many authors have developed methods to estimate some aspects of a site’s structure from the positions of neighboring sites, allowing the introduction of directional non-bonded terms to more accurately model directional interactions like hydrogen bonding [48–59]. Some protein models have estimated the orientation of hydrogen-bonding peptide groups from the positions of contiguous α -carbons [48–53] or from the positions of multiple backbone sites per residue [54], while others have estimated the orientation of ellipsoidal side-chain sites using the position of the neighboring backbone site [55]. Similar efforts have also been applied to DNA [56–58] and lipid [59] models.

In one protein example, recognizing that the orientation of peptide bond dipoles in the Protein Data Bank (PDB) [60] correlates with the angle among the three nearest α -carbons, Alemani *et al.* created a model defining the peptide bond dipole orientation as a function of the α -carbon angle [51]. By including a dipole interaction, a bistable bond angle interaction, and a term coupling consecutive dihedral angles along the α -carbon chain, Alemani *et al.* could tune between alpha-helix and beta-sheet secondary structures. Although this approach illustrates a minimal set of ingredients needed to select secondary structures, the low correlation in the PDB between peptide bond angles and backbone angles *within* each class of secondary structure (Fig. 1 in Ref. [51]) illustrates the resolution limit of such an approach.

A few authors have increased the resolution of coarse-grained models by introducing structure into individual coarse-grained sites in the form of vectors. Morriss-Andrews *et al.* created a nucleic acid model with ellipsoidal nucleobases allowed to freely rotate about the vector between the base site and its associated backbone site [61]. Spiga *et al.* created a protein model with electric dipoles at polar side-chain sites allowed to freely rotate in all directions [62]. Very recently, our group created a model for peptoids (positional isomers of peptides) assigning a rotatable vector to each site, using the vectors to construct orientation-dependent bonded and non-bonded interactions [63].

Here, we introduce a new method to bring atomistic resolution to coarse-grained models: assign an *independent three-dimensional orientation* to each site. While adding orientations to a point-site model increases the number of degrees of freedom per site from three to six, it efficiently uses the separation of energy scales characterizing organic chemistry: bond lengths and angles are stiff, while dihedral angles (except in rings and double bonds) are soft. Preferentially integrating out the stiff degrees of freedom allows a dramatic reduction in the number of degrees of freedom with a minimal loss of accuracy. As we discuss in Section 1, the ability of a coarse-grained model to accurately calculate observables depends on two contributions to its accuracy: (1) the accuracy of the effective interactions used to approximate the many-body potential of mean force and (2) the accuracy of atomic coordinates produced during a *roundtrip*

mapping. In a roundtrip mapping, atomic coordinates are mapped onto a coarse-grained model (throwing out information), then a new set of atomic coordinates are produced from the coarse-grained coordinates using analytical backmapping functions [64]. In Section 3, we explicitly demonstrate that an oriented coarse-grained protein model with a 4:1 mapping (eight atoms per oriented site) can achieve an unprecedented combination of efficiency and accuracy in a roundtrip mapping. Using all experimental protein structures from the Protein Data Bank [60] as a large and experimentally based proxy for an equilibrium distribution of atomic coordinates, we optimize backmapping functions to reduce the information lost during a roundtrip mapping. After this optimization, we find that heavy atoms move only 0.051 Å on average during the roundtrip mapping, while hydrogens move 0.179 Å.

Although the mapping between all-atom and coarse-grained representations is just one half of a coarse-grained model, our results bode well for developing tractable effective interactions that introduce a minimal amount of additional error. In contrast to point-site models, the position and orientation of a single site can predict the position of surrounding atoms associated with that site. Thus, the atomistic interactions between two sites can be expressed directly as a potential of mean force, calculated *without fitting* by recording the distribution of relative positions and orientations of the two sites in all-atom simulations. Although approximate factorizations and symmetries may simplify some of these calculations, in general we expect that these potentials of mean force will be six-dimensional (three positional and three orientational degrees of freedom). We leave the substantial computational task of calculating these potentials of mean force for future work.

1. Mapping and backmapping functions

Mapping and backmapping functions are essential to coarse-grained and multiscale modeling. Coarse-grained models consist of two parts, mapping functions and effective interactions. The set of forward mapping functions \mathbf{M} defines the set of coarse-grained coordinates \mathbf{R} via $\mathbf{R} = \mathbf{M}(\mathbf{r})$, where \mathbf{r} is the set of atomic positions. In the next paragraph we will discuss how backmapping functions $\mathbf{B}(\mathbf{R}) = \mathbf{r}_{\text{back}}$ allow coarse-grained models to approximately calculate even those observables that depend explicitly on the atomic positions \mathbf{r} . First, it is instructive to note that coarse-grained models can in principle be used to exactly calculate any observable that depends only on the reduced coordinates \mathbf{R} . For equilibrium observables, this can be seen by integrating out the variables eliminated by \mathbf{M} from the equilibrium distribution function [18, 19, 65]. This results in an expression for the effective interactions $V(\mathbf{R})$ that would ideally be applied

to the coarse-grained model:

$$V(\mathbf{R}) = -k_{\text{B}}T \ln \left(\frac{\Omega}{\omega} \int d\mathbf{r} \delta(\mathbf{M}(\mathbf{r}) - \mathbf{R}) \exp(-\beta v(\mathbf{r})) \right). \quad (1)$$

In Eq. 1, $v(\mathbf{r})$ is the all-atom potential energy function, and Ω and ω are the configurational volumes of the all-atom and coarse-grained systems. The exact solution to Eq. 1 includes many-body terms of all orders, which would be prohibitively difficult to calculate and prohibitively slow to simulate. Coarse-grained models are useful when Eq. 1 can be approximated by a small number of terms, e.g.

$$V(\mathbf{R}) \simeq \sum_{i=1}^{N_{\text{sites}}} \sum_{j=1}^{i-1} V_{\text{pair}}(\mathbf{R}_i, \mathbf{R}_j), \quad (2)$$

where V_{pair} is a pair interaction between sites \mathbf{R}_i and \mathbf{R}_j . We will show that backmapping functions can be developed to predict atomic positions from the position and orientation of individual coarse-grained sites, $\vec{b}_j = \vec{B}_j(\mathbf{R}_i)$. Thus, many of the atomic-scale pair interactions represented by Eq. 1 should be expressible by coarse-grained pair potentials of the form of Eq. 2. This will allow accurate effective interactions to be directly inverted from pair distribution functions calculated with small all-atom simulations, minimizing the need for fitting.

Backmapping functions are equally crucial in their own right, either for predicting observables that depend explicitly on the atomic coordinates \mathbf{r} or, as will be discussed in Section 4, for interfacing with all-atom simulations in multiscale schemes. Suppose that we want to use a coarse-grained model to calculate an observable \mathcal{O} that depends explicitly on the atomic positions \mathbf{r} . For simplicity, assume that \mathcal{O} depends linearly on \mathbf{r} , $\mathcal{O} = \mathbf{O} \cdot \mathbf{r}$. In addition to the error associated with approximating Eq. 1 by e.g. Eq. 2, we expect an error resulting from our inability to precisely know the positions \mathbf{r} . We can only calculate \mathcal{O} via a set of backmapping functions \mathbf{B} defining backmapped atomic positions $\mathbf{r}_{\text{back}} = \mathbf{B}(\mathbf{R})$. The error associated with this backmapping (beyond any error associated with the effective interactions) can be calculated by defining the roundtrip root-mean-square displacement (rmsd)

$$\Delta_{\mathbf{r}} = \left(\left\langle (\mathbf{r}_{\text{back}} - \mathbf{r})^2 \right\rangle \right)^{1/2} = \left(\left\langle (\mathbf{B}(\mathbf{M}(\mathbf{r})) - \mathbf{r})^2 \right\rangle \right)^{1/2}, \quad (3)$$

where the average is over the equilibrium distribution of the atomistic system. The error in \mathcal{O} depends on the covariance matrix for \mathbf{r} , $\Sigma_{\mathbf{r}}$, via $\Delta\mathcal{O} = (\mathbf{O} \cdot \Sigma_{\mathbf{r}} \cdot \mathbf{O}^{\text{T}})^{1/2}$, and the elements of $\Sigma_{\mathbf{r}}$ are bounded by the elements of $\Delta_{\mathbf{r}}$ by the Pearson correlation coefficient equation. Thus, to reduce the error in \mathcal{O} associated with backmapping, we want to reduce the roundtrip rmsd of all the atomic positions.

2. Method

Our method can be described generally as follows. We defined a set of forward mapping functions $\mathbf{M}(\mathbf{r}) = \mathbf{R}$ that would integrate out the stiff degrees of freedom found in organic chemistry (bond lengths and angles, double-bond dihedral angles, and dihedral angles in rings) while retaining a nearly 1:1 correspondence between \mathbf{M} and the soft degrees of freedom (non-ring, single-bond dihedral angles). Then, we defined backmapping functions $\mathbf{B}(\mathbf{R}) = \mathbf{r}_{\text{back}}$ that would take advantage of this 1:1 correspondence to back out atomic coordinates with a precision limited only by fluctuations in the stiff degrees of freedom. Finally, we numerically calculated optimal coefficients for the backmapping functions that minimized the root-mean-square displacement between initial and backmapped atomic positions ($\Delta_{\mathbf{r}}$ in Eq. 3), when applying the roundtrip mapping to a set of atomic configurations.

This procedure could be applied to any set of atomic configurations. The formal connection to a potential of mean force requires that the set be an equilibrium distribution of configurations. However, since there are no such experimental distributions with sufficient statistics to perform the optimization of \mathbf{B} , and since creating such a distribution with all-atom simulations would be computationally expensive, we instead used the Protein Data Bank (PDB) [60] as a large, convenient, and experimentally based proxy for an equilibrium distribution. In a general sense this is analogous to using the PDB to construct knowledge-based potentials for coarse-grained models [66]. However, in contrast to these approaches, our use of the PDB as a proxy only assumes that the local degrees of freedom integrated out by \mathbf{M} are in equilibrium in the PDB, making no assumptions about non-local degrees of freedom like those responsible for protein folding.

Details about how we extracted data from the PDB, dealt with missing or incorrect data, and treated indistinguishable atoms appear in the Supporting Information. The Supporting Information also contains the optimal parameters for the backmapping functions, a list of rmsds by amino acid and atom type, and the C code we used to optimize the backmapping functions, calculate the rmsds, and generate the molecular files and TCL scripts used to create the VMD [67] images in this paper.

3. Results

Figure 1 illustrates how we constructed forward mapping and backmapping functions to take advantage of the stiff degrees of freedom found in proteins. As shown in Fig. 1 (a) for one non-terminal glutamine residue drawn from the PDB, we grouped each amino acid residue into one to three coarse-grained sites \mathbf{r}_i , each defined by a position and three orthonormal vectors, $\mathbf{R}_i = \{\vec{R}_i, \hat{E}_{ix}, \hat{E}_{iy}, \hat{E}_{iz}\}$. We defined the forward map-

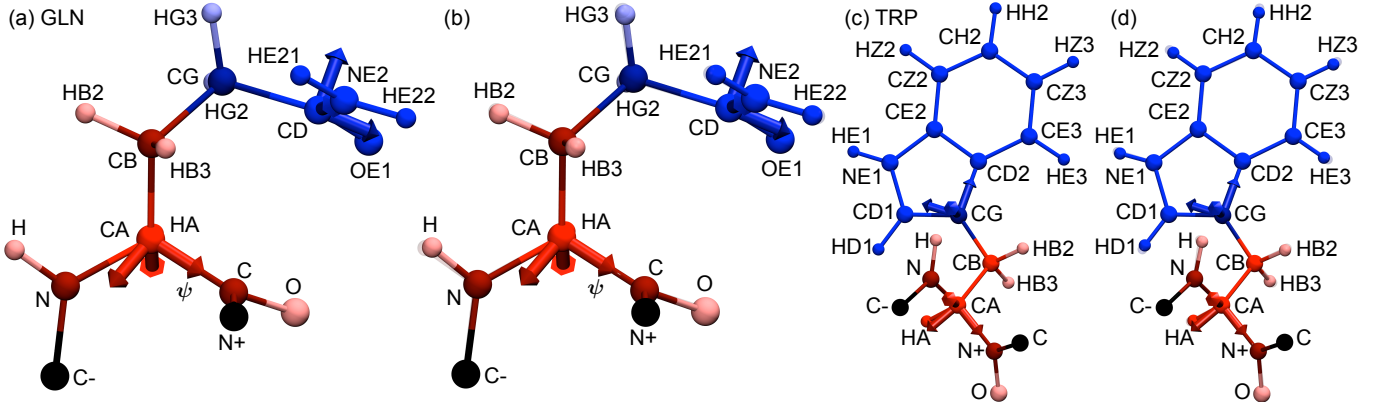


FIG. 1: Illustration of forward mapping and backmapping functions for non-terminal glutamine and tryptophan residues, using example residues from the PDB structure 2iqx of malate synthase G [68]. Atoms are labeled using standard PDB atom names. Examples for the other eighteen amino acids appear in Fig. 2 and the Supporting Information. (a) Glutamine atoms from the PDB structure (spheres) are mapped to two oriented sites (vectors). The backbone site (red) is defined by the positions of the atoms CA, C, and N, and the sidechain site (blue) is defined by the atoms CD, NE2, and OE1. Atoms are colored according to to which site they are associated with. Neighboring atoms in adjacent residues are colored black. (b) Backmapped atom positions (opaque) superposed on original positions (semi-transparent). Lighter-colored atoms (e.g. HB2 and HB3) are backmapped using a linear correction that depends on the backmapped position of the nearest atom in an adjacent site (e.g. CG), which is colored darker. (c) A tryptophan residue from 2iqx is mapped onto two oriented sites. The backbone site (red) is defined by the positions of the atoms CA, C, and N, and the sidechain site (blue) is defined by the atoms CG, ND2, and CD1. (d) Backmapped atom positions (opaque) superposed on original positions, as in (b).

ping function \mathbf{M}_i for each site i as a function of the positions of three atoms, $\vec{r}_{\nu_{i1}}$, $\vec{r}_{\nu_{i2}}$, and $\vec{r}_{\nu_{i3}}$. (ν_{in} defines the index of the n th atom defining the i th site.) We let $\vec{r}_{\nu_{i1}}$ define the site position, $\vec{R}_i = \vec{r}_{\nu_{i1}}$, and we let the other two sites define the orthonormal vectors,

$$\begin{aligned}\hat{E}_{ix} &= \hat{N}(\vec{r}_{\nu_{i1}} - \vec{r}_{\nu_{i0}}), \\ \hat{E}_{iz} &= \hat{N}((\vec{r}_{\nu_{i1}} - \vec{r}_{\nu_{i0}}) \times (\vec{r}_{\nu_{i2}} - \vec{r}_{\nu_{i0}})), \\ \hat{E}_{iy} &= \hat{N}(\hat{E}_{iz} \times \hat{E}_{ix}),\end{aligned}\quad (4)$$

where $\hat{N}(\vec{r}) = \vec{r}/|\vec{r}|$. For example, we placed a backbone site (red arrows in Fig. 1 (a)) at each α -carbon (CA) and defined the backbone site orientation by the positions of the neighboring carbonyl carbon (C) and nitrogen (N) atoms in the backbone. Since bond lengths and angles tend to be distributed close to their average values, we could immediately write down backmapping functions for all atoms directly bound to the central atom. For an atom j of type α (e.g. β -carbons in glutamine) belonging to site i , we simply defined the backmapping function $\vec{B}_\alpha(\mathbf{R}_i)$ via

$$\vec{B}_\alpha(\mathbf{R}_i) = \vec{R}_i + c_{\alpha x}\hat{E}_{ix} + c_{\alpha y}\hat{E}_{iy} + c_{\alpha z}\hat{E}_{iz}, \quad (5)$$

where \vec{c}_α is the average position of atoms of type α in the frame of their coarse-grained site. With this backmapping, the roundtrip rmsd becomes simply the rmsd of the atom positions in the frame of the site. As illustrated by the close agreement between backmapped (solid) and original (semi-transparent) atom positions in Fig. 1 (b), all atoms bound to the α -carbon tend to move only small distances during the roundtrip mapping. Averaging over

all residues in the PDB, we find that α -carbons, carbonyl carbons, nitrogens, β -carbons (CB), and α -hydrogens (HA) move on average 0, 0.01, 0.07, 0.08, and 0.06 Å, respectively. We can account for the model's ability to backmap five atom positions (15 degrees of freedom) from one site (six degrees of freedom) because the tetrahedral cluster is constrained by four bond lengths and five independent bond angles, each largely conserved across the PDB.

We found that directly backmapping backbone atoms separated by two bonds from the α -carbon did not result in low roundtrip rmsds, because the positions of these atoms in the frame of the backbone site each depend on a soft torsional degrees of freedom. For example, the position of the carbonyl oxygen (O) in the i th residue depends on the ψ dihedral angle (torsion of the CA-C bond, see Fig. 1 (a)). Fortunately, the same torsional degree of freedom controls the position of the nitrogen in the $(i+1)$ th residue, denoted N^+ in Fig. 1. Since the position of the $(i+1)$ th nitrogen is predicted accurately by direct backmapping from the $(i+1)$ th backbone site, we could predict the position of the i th oxygen by fitting it to a linear function of the predicted position of the $(i+1)$ th nitrogen. All of this is done in the frame of the i th backbone site. A linear function is sufficient because the dihedral angle rotation relating the oxygen and nitrogen positions is a linear transformation. In general, for an atom l of type β directly backmapped by site k , we define $\vec{B}'_\alpha(\mathbf{R}_i, \mathbf{R}_k)$ as its predicted position in the frame of site i ,

$$\vec{B}'_\beta(\mathbf{R}_i, \mathbf{R}_k) = (\vec{B}_\beta(\mathbf{R}_k) - \vec{R}_i) \cdot \{\vec{E}_{ix}, \vec{E}_{iy}, \vec{E}_{iz}\}. \quad (6)$$

Res	Atom 1	Atom 2	Atom 3	Direct atoms	Corrected atoms
ALA	CA	C	N	HA CB HB2 HB3 HB1	O H
ARG	CA	C	N	HA	O H
ARG	CG	CD	CB	HG2 HG3	HB2 HB3 HD2 HD3
ARG	CZ	NH1	NH2	NE HH11 HH12 HH21 HH22	HE
ASN	CA	C	N	HA CB	O H HB2 HB3
ASN	CG	ND2	OD1	HD21 HD22	
ASP	CA	C	N	HA CB	O H HB2 HB3
ASP	CG	OD2	OD1	HD2	
CYS	CA	C	N	HA	O H
CYS	SG	CB	HG		HB2 HB3
GLN	CA	C	N	HA CB	O H HB2 HB3
GLN	CD	NE2	OE1	CG HE21 HE22	HG2 HG3
GLU	CA	C	N	HA CB	O H HB2 HB3
GLU	CD	OE2	OE1	CG HE2	HG2 HG3
GLY	CA	C	N	HA2 HA3	O H
HIS	CA	C	N	HA CB	O H HB2 HB3
HIS	CG	ND1	CD2	CE1 NE2 HD1 HD2 HE1 HE2	
ILE	CA	C	N	HA	O H
ILE	CG1	CB	CD1	HG12 HG13 HD11 HD12 HD13	CG2 HG21 HG22 HG23 HB
LEU	CA	C	N	HA CB	O H HB2 HB3
LEU	CG	CD1	CD2	HG HD11 HD12 HD13 HD21 HD22 HD23	
LYS	CA	C	N	HA	O H
LYS	CG	CB	CD	HG2 HG3	HB2 HB3
LYS	CE	NZ	CD	HE2 HE3 HZ1 HZ2 HZ3	HD2 HD3
MET	CA	C	N	HA CB	O H HB2 HB3
MET	SD	CE	CG	HE1 HE2 HE3	HG2 HG3
PHE	CA	C	N	HA CB	O H HB2 HB3
PHE	CG	CD1	CD2	CE1 CE2 CZ HD1 HD2 HE1 HE2 HZ	
PRO	CA	C	N	HA	O H
PRO	CG	CD	CB	HG2 HG3	HB2 HB3 HD2 HD3
SER	CA	C	N	HA	O H
SER	OG	CB	HG		HB2 HB3
THR	CA	C	N	HA	O H
THR	OG1	CB	HG1		HB CG2 HG21 HG22 HG23
TRP	CA	C	N	HA CB	O H HB2 HB3
TRP	CG	CD2	CD1	NE1 CE2 CE3 CZ2 CZ3 CH2 HD1 HE1 HE3 HZ2 HZ3 HH2	
TYR	CA	C	N	HA CB	O H HB3 HB2
TYR	CG	CD1	CD2	CE1 CE2 CZ OH HD1 HD2 HE1 HE2 HH	
VAL	CA	C	N	HA	O H
VAL	CB	CG1	CG2	HB HG11 HG12 HG13 HG21 HG22 HG23	

TABLE 1: List of atoms defining the forward mapping and backmapping functions for each site. The first column lists the residue, the next three columns list the atoms ν_{ij} , $j = 1, 2, 3$, defining the forward mapping function \mathbf{M}_i . The fifth column lists additional atoms backmapped directly by Eq. 5. The last column lists the atoms backmapped by Eq. 7.

Then, if j is the atom of type α in site i whose position is correlated with the position of atom l of type β , the backmapping function for j with a linear correction is

$$\vec{B}_\alpha(\mathbf{R}_i, \mathbf{R}_k) = \vec{R}_i + c'_{\alpha\beta x} \hat{E}_{ix} + c'_{\alpha\beta y} \hat{E}_{iy} + c'_{\alpha\beta z} \hat{E}_{iz}, \quad (7)$$

where

$$\vec{c}'_{\alpha\beta} = \vec{I}_{\alpha\beta} + S_{\alpha\beta} \cdot \vec{B}'_\beta(\mathbf{R}_i, \mathbf{R}_k). \quad (8)$$

We calculated the 3-vector $\vec{I}_{\alpha\beta}$ and the 3×3 matrix $S_{\alpha\beta}$ by analytically performing ordinary least-squares fits to the three components of the vector equation

$$\left\langle (\vec{r}_j - \vec{R}_i) \cdot \{\vec{E}_{ix}, \vec{E}_{iy}, \vec{E}_{iz}\} \right\rangle_{j \in \alpha} = \vec{c}'_{\alpha\beta}, \quad (9)$$

where the average is over all atoms j of type α .

The three-bond spacing between α -carbons in proteins is ideal for backmapping accurate atomic positions from oriented backbone sites. For each residue we backmapped the C, N, HA, and CB atoms directly via Eq. 5, and we backmapped the O and H atoms via Eq. 7 using the predicted positions of the N^+ and C^- atoms, respectively. As illustrated by the close agreement between the solid and semi-transparent O and H atoms in Fig. 1 (b), the linear correction resulted in small displacements during the roundtrip mapping. Averaged over the PDB, O atoms moved 0.086 Å and H atoms moved 0.274 Å. If residues were separated by more than three bonds, such an accurate backmapping would not be possible with only one backbone site per residue, because additional soft

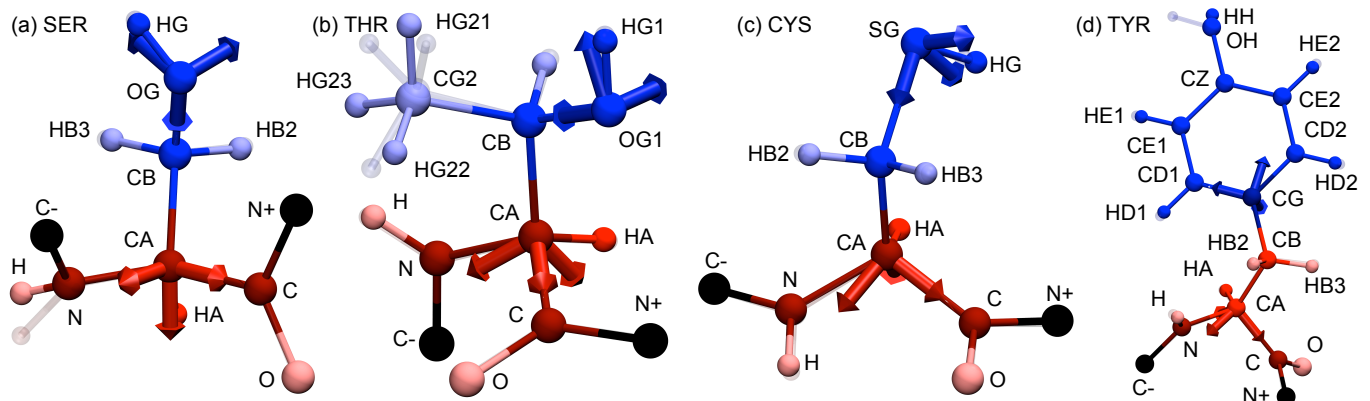


FIG. 2: Backmapped atom positions (opaque) superposed on original positions (semi-transparent) for (a) serine, (b) threonine, (c) cysteine, and (d) tyrosine. In (a-c) the hydroxyl or thiol groups help define the sidechain site. In (d) the torsion of the hydroxyl group is unconstrained, leading to a large disagreement between the original and backmapped HH position.

torsional degrees of freedom would be unconstrained.

Moving up the main branch of each amino acid sidechain, we added as few additional sites as possible while ensuring that sites were spaced by no more than three bonds. Columns 2-4 of Table 1 lists the atoms defining each site, using standard atom names from the PDB. As for the backbone sites, we used Eq. 5 to directly backmap those atoms whose relative positions could be constrained by stiff degrees of freedom (Column 5 of Table 1), and we used linear corrections to backmap those atoms off the main chain that required a torsional degree of freedom to be constrained (last column of Table 1). When correcting the position of an atom, we always corrected it using the predicted position of the closest main-chain atom on a neighboring site. For example, for glutamine (Fig. 1(a)) we corrected the β -hydrogens in the backbone site (HB2 and HB3) by the predicted position of the γ -carbon in the sidechain site (CG), and we corrected the γ -hydrogens in the sidechain site (HG2 and HG3) by the predicted position of the β -carbon in the backbone site (CB). The close agreement between solid and semi-transparent HB2, HB3, HG2, and HG3 atoms in Fig. 1 (b) illustrates the high accuracy of these roundtrip mappings.

Although the sites in some sidechains had to be spaced by fewer than three bonds, using oriented sites more than made up for this redundancy by (1) utilizing the stiff torsional degrees of freedom in the rings of histidine, phenylalanine, tryptophan, and tyrosine and (2) allowing accurate estimation of methyl and amine hydrogens with no additional overhead. Tryptophan’s double ring is a dramatic example of the first case. As shown in Fig. 1 (c), we placed tryptophan’s sidechain site on the γ -carbon (CG), only two bonds away from the α -carbon (CA) of the backbone site. This meant that the β -carbon (CB) could be predicted by either site; we chose to associate it with the backbone. Although this meant that the sidechain site’s orientation was not used to predict the position of the β -carbon, the stiffness of the double ring meant that we

could use the site’s orientation to predict the position of *twelve* atoms not directly bonded to the γ -carbon. As illustrated in Fig. 1 (d), the direct backmapping of these atoms resulted in low roundtrip rmsds. Averaging over the PDB, the largest roundtrip displacement in tryptophan’s double ring was for HZ3, which moved on average 0.13 Å. Our ability to model ring groups using few degrees of freedom starkly contrasts the ability of point-size coarse-grained models that typically require more sites per atom to model rings [69, 70].

Glutamine’s amine group (Fig. 1 (c)) is an example of the second case, where methyl or amine hydrogens could be backmapped with high accuracy and no additional overhead. When adding sites, we ensured that methyl carbons and amine nitrogens were within one bond of a coarse-grained site, but we imposed no such requirement for the methyl and amine hydrogens. As a result, the hydrogens’ positions in the frame of the site depend on the torsion of the bond connecting the carbon or nitrogen to the site center. However, we found that these hydrogen positions could be backmapped directly via Eq. 5 with relatively low rmsds of 0.11 Å for neutral amine hydrogens in arginine, asparagine, and glutamine; between 0.17 Å and 0.24 Å for methyl hydrogens in alanine, isoleucine, leucine, methionine, threonine, and valine; and 0.25 Å for quaternary ammonium hydrogens in lysine. The low rmsds for the neutral amine hydrogens is due to the planarity of the guanidinium and amide groups. The rmsds of the methyl and NH_3^+ groups are kept reasonably low from a combination of two factors. First, we named the indistinguishable hydrogen atoms according to their dihedral angles, which imposes an upper bound of around 0.51 Å for methyl groups and 0.52 Å for NH_3^+ groups (see Supporting Information). Second, these groups do show some preference for certain dihedral angles, reducing the rmsd to less than half of these upper bounds.

In contrast to the amine and methyl hydrogens, we found that hydroxyl and thiol groups in serine, thre-

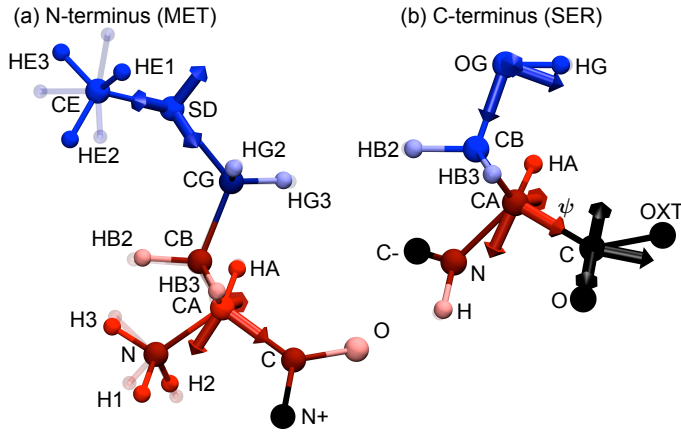


FIG. 3: Backmapped atom positions (opaque) superposed on original positions (semi-transparent) for (a) the N-terminal methionine and (b) the C-terminal serine from the PDB structure 2eyz of the CRK protein [71]. In (a) the N-terminal quaternary ammonium hydrogens are backmapped from the position and orientation of the C_α site via Eq. 5. In (b) an additional C-terminal site is defined from the positions of atoms C, O, and OXT.

onine, tyrosine, and cysteine have broad distributions of dihedral angles, and the single terminal hydrogens in these groups have no indistinguishable partners that could be used to reduce their rmsd. For serine, threonine, and cysteine, we were able to include the hydrogen as one of the three atoms defining the site without increasing the number of sites or adversely affecting the backmapping of other atoms, as shown in Fig. 2 (a-c). For tyrosine, the hydroxyl group lies on the opposite side of an aromatic ring from the site used to backmap the ring. We chose to leave the hydroxyl torsion unconstrained in the backmapping, leading to an unusually large rmsd of 0.836 Å for the hydroxyl hydrogen atom (HH). Since introducing another site (six degrees of freedom) to constrain this one degree of freedom would be inefficient, the best way to reduce the rmsd might be to introduce an additional internal degree of freedom to tyrosine’s sidechain site. This degree of freedom would control the backmapping of HH and modulate the site’s effective interactions.

We modeled the N-terminal hydrogens similarly to how we modeled the amine hydrogens on the sidechains, backmapping their positions directly via Eq. 5 using the terminal backbone site, as shown in Fig. 3 (a). Since an appreciable number (2.4%) of N-terminal amines in the PDB were neutral, we treated NH_2 and NH_3^+ groups separately. Although these terminal amine groups occupy a broad distribution of dihedral angles, we were able to reduce the rmsd by naming the indistinguishable hydrogens according to their dihedral angles, bringing the rmsds down to 0.451 Å for NH_2 hydrogens and 0.637 Å for NH_3 hydrogens (see Supporting Information). As for tyrosine’s hydroxyl group, introducing a single internal degree of freedom to describe this torsion could significantly reduce these rmsds.

Residue	Atoms	Sites	Mapping	Heavy rmsd	H rmsd
ALA	10	1	5.0	0.056	0.178
ARG	24	3	4.0	0.043	0.120
ASN	14	2	3.5	0.054	0.158
ASP	12	2	3.0	0.053	0.190
CYS	11	2	2.8	0.058	0.163
GLN	17	2	4.2	0.050	0.141
GLU	15	2	3.8	0.052	0.160
GLY	7	1	3.5	0.059	0.358
HIS	18	2	4.5	0.050	0.140
ILE	19	2	4.8	0.050	0.160
LEU	19	2	4.8	0.049	0.159
LYS	22	3	3.7	0.046	0.163
MET	17	2	4.2	0.078	0.191
PHE	20	2	5.0	0.048	0.135
PRO	15	2	3.8	0.050	0.063
SER	11	2	2.8	0.057	0.131
THR	14	2	3.5	0.057	0.168
TRP	24	2	6.0	0.050	0.134
TYR	21	2	5.2	0.053	0.334
VAL	16	2	4.0	0.045	0.209
Total				0.051	0.179

TABLE 2: List of roundtrip mapping ratios and roundtrip rmsds by amino acid (top 20 rows) and overall (bottom row). The mapping ratio R is the ratio of the number of atomic degrees of freedom to the number of coarse-grained degrees of freedom, $R = N_{\text{atoms}}/(2 \times N_{\text{sites}})$. The rmsds are separated for heavy atoms and hydrogens.

As shown in Fig. 3 (b), the positions of the heavy atoms O and OXT at the C-terminus depend on the torsion of the C-terminal $CA-C$ bond (ψ dihedral angle). To allow accurate backmapping, we introduced an additional site at the C-terminus defined by the positions of the atoms C, O, and OXT (black arrows in Fig. 3 (b)). Even with this additional site, the rmsd for the OXT atom is still moderately large, 0.347 Å, mostly due to an unusually broad distribution for the C-OXT bond length in the PDB, $r_{C-OXT} = 1.249 \pm 0.327$ Å.

We designed our model to target proteins in their most common charged states, but we also applied our roundtrip mapping to structures containing protonated carboxyl groups at the C-terminus and/or in aspartic acid or glutamic acid sidechains, which represented fewer than 1% of the 2% of PDB structures containing hydrogens. Since the positions of the hydrogens in these groups depend on an unconstrained torsional degree of freedom, they had predictably large rmsds of 0.973 Å, 0.809 Å, and 0.804 Å, respectively. For the C-terminus and aspartic acid, these hydrogens could be constrained by using an alternate mapping for charged groups. For example, changing the atoms defining the C-terminus site from C, O, and OXT (see Fig. 3 (b)) to OXT, HXT, and C would constrain HXT up to variations in bond lengths and angles, while allowing O to be backmapped through Eq. 7. The main chain of the glutamic acid sidechain is too long for the additional hydrogen to be accurately backmapped without adding an additional degree of freedom.

Overall, we found that using oriented coarse-grained

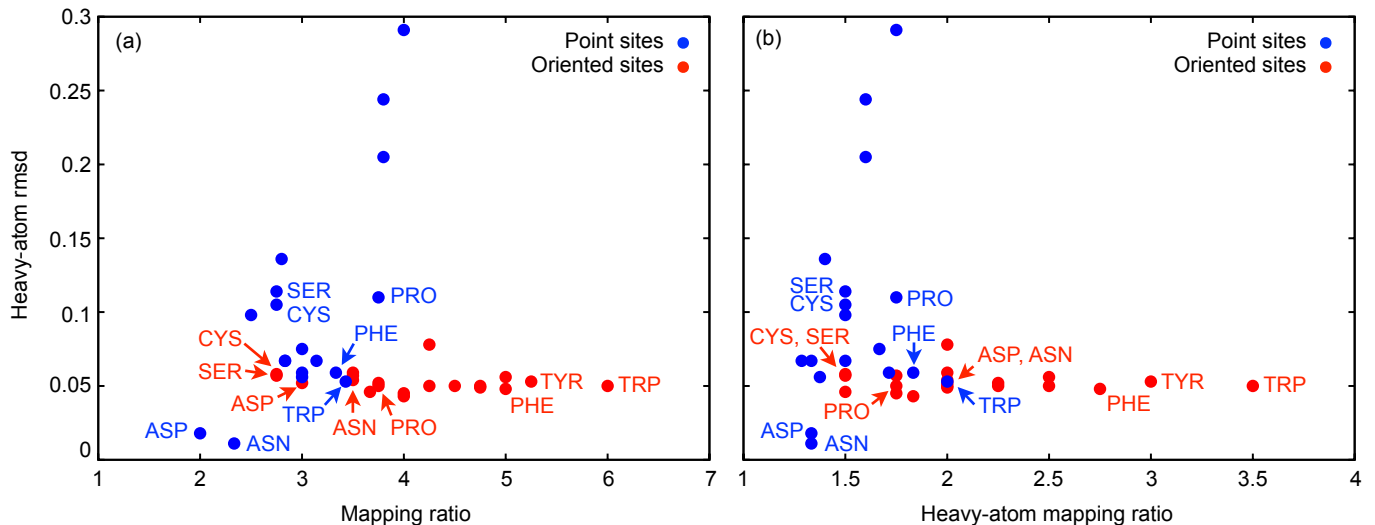


FIG. 4: Comparison of heavy-atom rmsds and mapping ratios for our model (red points) and an “intermediate-resolution” point-site model (PRIMO [54, 64]) specifically designed for accurate backmapping (blue points). Each point represents a different amino acid. Panel (a) defines the mapping ratio as the number of atomic degrees of freedom per coarse-grained degree of freedom. Panel (b) defines the mapping ratio as the number of non-hydrogen atomic degrees of freedom per coarse-grained degree of freedom. Results for the point-site model are taken from Ref. [64] and were obtained by performing a roundtrip mapping on 611 protein structures minimized with an all-atom forcefield. In general, our model is both more accurate (smaller rmsds) and more efficient (larger mapping ratios).

sites allowed us to efficiently store the information necessary to backmap accurate all-atom configurations. Averaged over the PDB, heavy atoms moved 0.051 Å during our roundtrip mapping, while hydrogens moved 0.179 Å. Table 2 lists the mapping ratios and rmsd for each amino acid. A complete list broken down by atom type appears in the Supporting Information. The low rmsds are remarkable given that the average mapping ratio is 4.1; that is, information associated with 76% of the atomic degrees of freedom are discarded during the roundtrip mapping.

To get a sense of how much of this efficiency can be attributed to using oriented sites, in Fig. 4 (a) we compare heavy-atom rmsds and mapping ratios for our model with those calculated in Ref. [64] for an “intermediate-resolution” point-site model (PRIMO) which was designed for accurate backmapping [64]. PRIMO’s high resolution [64] and detailed effective interactions [54] has allowed it model folded, folding [54], and membrane [72] proteins with higher accuracy than typical for coarse-grained models. PRIMO maps each residue to between 4 and 8 structureless sites. As a result, it reduces the number of atomic degrees of freedom by factors ranging between 2 (aspartic acid) and 4 (valine), as shown by the blue points in Fig. 4 (a), with an average mapping ratio of 3.0. This is a substantially finer mapping than most point-site coarse grained models. The average mapping ratios for the UNRES [73], MARTINI [70], and MS-CG [74] models are 8.2, 7.1, and 6.0, while Spiga *et al.*’s model including dipole degrees of freedom has an average mapping ratio of 6.6. Ref. [64] only performed a roundtrip mapping with PRIMO for heavy atoms (find-

ing an overall rmsd of 0.099 Å), so we only plot heavy-atom rmsds in Fig. 4. Since the calculation of the heavy-atom rmsds does not use the mapping functions for hydrogens, it is instructive to also compare the heavy-atom rmsds to *heavy-atom* mapping ratios, the number of non-hydrogen atomic degrees of freedom per coarse-grained degree of freedom. As shown by the blue points in Fig. 4 (b), PRIMO’s heavy-atom mapping ratios fall between 1.3 (lysine) and 2 (tryptophan), with an average of 1.5.

Comparing PRIMO to our model (red points in Fig. 4), we find that across amino acids our model is more accurate (smaller rmsds) and more efficient (larger mapping ratios). The worst cases for our model are cysteine and serine, whose sidechains are two heavy atoms long, requiring a sidechain site to backmap the position of only one heavy atom (and three hydrogens) that could not be directly backmapped from the backbone site (see Fig. 2 (a) and (c)). Our mapping ratios for these amino acids are no larger than PRIMO’s (2.75, or 1.5 considering only heavy atoms), but our backmapping is more accurate (rmsds of 0.58 and 0.57 Å vs 0.105 and 0.114 Å for PRIMO). Our proline mapping also has a ratio equal to its PRIMO counterpart (3.75, or 1.75 considering only heavy atoms) and a lower rmsd (0.050 Å vs 0.110 Å). The only residues with lower rmsds in PRIMO are asparagine, aspartic acid, and methionine (0.011, 0.018, and 0.067 Å vs 0.054, 0.053, and 0.078 Å), but PRIMO achieves these low rmsds by using a nearly atomistic heavy-atom mapping ratio of 4/3, compared to a ratio of 2 used for these amino acids in our model. The remaining fourteen amino acids have both larger mapping ratios and smaller rmsds in our model. Amino acids with

rings perform best in our model, due to the fact that the planarity of the rings allows the structure of an entire ring (or double ring) to be accurately stored in a single oriented site. Histidine, phenylalanine, tyrosine, and tryptophan all have smaller rmsds than in PRIMO despite having mapping ratios that are at least 50% larger.

4. Discussion

By demonstrating that our model can store atomic positions with an unprecedented combination of accuracy (small rmsds) and efficiency (large mapping ratios), we have justified the substantial computational task of parameterizing the effective interactions necessary to complete our model. Calculating multidimensional potentials of mean force from all-atom simulations should allow us to write down effective interactions capturing atomic-scale interactions, because the relative atomic positions controlling these interactions are related directly to the relative positions and orientations of neighboring coarse-grained sites. As discussed in Section 1, equipping our model with these effective interactions should allow us to calculate any observable of the atomistic system with an error proportional to the small roundtrip rmsds calculated in this paper.

Having designed a model that preferentially integrates out the stiffest degrees of freedom (bond bending, bond stretching, and dihedral rotations in rings) we think that we have approached the accuracy limit of coarse-grained modeling. Nevertheless, there are undoubtedly some molecular processes so sensitive to error that they could not be modeled even by an optimal coarse-grained model. For such systems, coarse-grained models can be used to accelerate equilibration, sampling, or dynamics of all-atom simulations in various multiscale schemes [21–33].

A small roundtrip rmsd is a clear figure of merit for a coarse-grained model’s performance in all multiscale schemes. Sequential multiscale modeling [21–23] requires that configurations be handed back and forth between coarse-grained and all-atom simulations without lengthy relaxations. Reducing the roundtrip rmsd ensures that relaxations are short. Embedded multiscale modeling [24–29] requires that coarse-grained and all-atom regions can be stitched together in the same simulation box. When a coarse-grained can backmap accurate all-atom configurations, the coarse-grained region can be seamlessly blended into the all-atom region using the backmapped all-atom configurations.

Multiscale replica exchange simulations [30–33] require that coordinates be exchanged in equilibrium along a ladder of otherwise equivalent system “replicas” differing in resolution. These exchanges allow the higher replicas (coarse-grained simulations) to accelerate sampling of the lowest replica (the all-atom simulation) without biasing the all-atom simulation. The efficiency of a

multiscale replica exchange simulation is determined by the number of replicas and the exchange acceptance rates. Using a high-resolution coarse-grained model in this approach would increase the overlap in distributions sampled by the highest and lowest replicas, thereby increasing acceptance rates and/or requiring fewer replicas.

5. Conclusion

We have demonstrated the accuracy of a coarse-grained protein model with oriented coarse-grained sites by calculating the distance atoms move during a roundtrip mapping. By preferentially integrating out stiff degrees of freedom associated with bond stretching, bond bending, and bond twisting in rings, our model achieves a combination of accuracy (small rmsds) and efficiency (large mapping ratios) that has not previously been attained by coarse-grained protein models. Our model’s nearly atomistic resolution should allow for parameterization of detailed effective interactions accounting for atomic-scale interactions. Once equipped with these effective interactions, our model should be able to extend the reach of time and length scales accessible to molecular simulation, either through direct simulations or through seamless integration into multiscale simulation schemes.

6. Acknowledgement

We thank Ranjan Mannige, Steve Whitelam, and Ron Zuckermann for useful comments on the manuscript. This project was funded by the Defense Threat Reduction Agency under Contract No. IACRO-B1144571. Work at the Molecular Foundry was supported by the Office of Science, Office of Basic Energy Sciences, of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

7. Supporting information

(1) Supporting document including (a) descriptions of how we extracted data from the PDB, dealt with missing or incorrect data, and treated indistinguishable atoms, (b) images illustrating the forward mapping and backmapping of the remaining amino acids, (c) list of roundtrip rmsds by amino acid and atom type, and (d) list of the calculated optimal parameters for the model. (2) Source code in C that we used to optimize the backmapping functions, calculate the rmsd, and generate the molecular files and TCL scripts used to create the VMD [67] images in this paper.

- [1] D. E. Shaw, P. Maragakis, K. Lindorff-Larsen, S. Piana, R. O. Dror, M. P. Eastwood, J. A. Bank, J. M. Jumper, J. K. Salmon, Y. Shan, et al., *Science* **330**, 341 (2010).
- [2] K. Lindorff-Larsen, S. Piana, R. O. Dror, and D. E. Shaw, *Science* **334**, 517 (2011).
- [3] S. Piana, K. Lindorff-Larsen, and D. E. Shaw, *J. Phys. Chem. B* **117**, 12935 (2013).
- [4] S. Piana, K. Lindorff-Larsen, and D. E. Shaw, *Proc. Natl. Acad. Sci. U.S.A.* **110**, 5915 (2012).
- [5] K. A. Feenstra, B. Hess, and H. J. C. Berendsen, *J. Comput. Chem.* **20**, 786 (1999).
- [6] M. Nuclea, R. Kaye, S. Milton, and C. G. Glabe, *J. Biol. Chem.* **14**, 10311 (2007).
- [7] M. Bartolini, M. Naldi, J. Fiori, F. Valle, F. Biscarini, D. V. Nicolau, and V. Andrisano, *Anal. Biochem.* **414**, 215 (2011).
- [8] P. Nguyen and P. Derreumaux, *Acc. Chem. Res.* **47**, 603 (2014).
- [9] S. O. Nielson, C. F. Lopez, G. Srinivas, and M. L. Klein, *J. Phys. Condens. Matter* **16**, R481 (2004).
- [10] C. Clementi, *Curr. Opin. Struct. Biol.* **18**, 10 (2008).
- [11] T. Murtola, A. Bunker, I. Vattulainen, M. Deserno, and M. Karttunen, *Phys. Chem. Chem. Phys.* **11**, 1869 (2009).
- [12] V. Tozzini, *Q. Rev. Biophys.* **43**, 3 (2010).
- [13] J. Trylska, *J. Phys. Condens. Matter* **22**, 453101 (2010).
- [14] S. C. L. Kamerlin, S. Vicatos, A. Dryga, and A. Warshel, *Annu. Rev. Phys. Chem.* **62**, 41 (2011).
- [15] C. Hyeon and D. Thirumalai, *Nat. Commun.* **2**, 487 (2011).
- [16] S. Takada, *Curr. Opin. Struct. Biol.* **22**, 130 (2012).
- [17] W. Shinoda, R. DeVane, and M. L. Klein, *Curr. Opin. Struct. Biol.* **22**, 175 (2012).
- [18] M. G. Saunders and G. A. Voth, *Annu. Rev. Biophys.* **42**, 73 (2013).
- [19] W. G. Noid, *J. Chem. Phys.* **139**, 090901 (2013).
- [20] E. Brini, E. A. Algaer, P. Ganguly, C. Li, F. Rodriguez-Ropero, and N. F. A. van der Vegt, *Soft Matter* **9**, 2108 (2013).
- [21] W. Tschöp, K. Kremer, O. Hahn, J. Batoulis, and T. Bürger, *Acta Polymer.* **49**, 75 (1998).
- [22] A. Y. Shih, P. L. Freddolino, S. G. Sligar, and K. Schulten, *Nano Lett.* **7**, 1692 (2007).
- [23] J. D. Perlmutter and J. N. Sachs, *Biochim. Biophys. Acta* **1788**, 2284 (2009).
- [24] M. Neri, C. Anselmi, M. Cascella, A. Maritan, and P. Carloni, *Phys. Rev. Lett.* **95**, 218102 (2005).
- [25] Q. Shi, S. Izvekov, and G. A. Voth, *J. Phys. Chem. B* **110**, 15045 (2006).
- [26] M. R. Machado, P. D. Dans, and S. Pantano, *Phys. Chem. Chem. Phys.* **13**, 18134 (2011).
- [27] A. B. Mamonov, S. Lettieri, Y. Ding, J. L. Sarver, R. Palli, T. F. Cunningham, S. Saxena, and D. M. Zuckerman, *J. Chem. Theory Comput.* **8**, 2921 (2012).
- [28] N. di Pasquale, D. Marchisio, and P. Carbone, *J. Chem. Phys.* **137**, 164111 (2012).
- [29] M. Leguèbe, C. Nguyen, L. Capece, Z. Hoang, A. Giorgetti, and P. Carloni, *PLOS ONE* **7**, e47332 (2012).
- [30] E. Lyman, F. M. Ytreberg, and D. M. Zuckerman, *Phys. Rev. Lett.* **96**, 028105 (2006).
- [31] E. Lyman and D. M. Zuckerman, *J. Chem. Theory Comput.* **2**, 656 (2006).
- [32] M. Christen and W. F. van Gunsteren, *J. Chem. Phys.* **124**, 154106 (2006).
- [33] K. Moritsugu, T. Terada, and A. Kidera, *J. Chem. Phys.* **133**, 244105 (2010).
- [34] A. A. Louis, *J. Phys. Condens. Matter* **14**, 9187 (2002).
- [35] F. H. Stillinger, H. Sakai, and S. Torquato, *J. Chem. Phys.* **117**, 288 (2002).
- [36] M. E. Johnson, T. Head-Gordon, and A. A. Louis, *J. Chem. Phys.* **126**, 144509 (2007).
- [37] P. Kowalczyk, P. A. Gauden, and A. Ciach, *J. Phys. Chem. B* **113**, 12988 (2009).
- [38] P. Kowalczyk, P. A. Gauden, and A. Ciach, *J. Phys. Chem. B* **115**, 6985 (2011).
- [39] D. Reith, M. Putz, and F. Müller-Plathe, *J. Comput. Chem.* **24**, 1624 (2003).
- [40] S. Izvekov and G. A. Voth, *J. Phys. Chem. B* **109**, 2469 (2005).
- [41] W. G. Noid, J.-W. Chu, G. S. Ayton, V. Krishna, S. Izvekov, G. A. Voth, A. Das, and H. C. Andersen, *J. Chem. Phys.* **128**, 244114 (2008).
- [42] W. G. Noid, P. Liu, Y. Wang, J.-W. Chu, G. S. Ayton, S. Izvekov, H. C. Andersen, and G. A. Voth, *J. Chem. Phys.* **128**, 244115 (2008).
- [43] S. Izvekov, P. W. Chung, and B. M. Rice, *J. Chem. Phys.* **133**, 064109 (2010).
- [44] M. S. Shell, *J. Chem. Phys.* **129**, 144108 (2008).
- [45] A. Chaimovich and M. S. Shell, *Phys. Rev. E* **81**, 060104 (2010).
- [46] A. Chaimovich and M. S. Shell, *J. Chem. Phys.* **134**, 094112 (2011).
- [47] S. P. Carmichael and M. S. Shell, *J. Phys. Chem. B* **116**, 8383 (2012).
- [48] A. Liwo, S. Oldziej, C. Czaplewski, U. Kozłowska, and H. A. Scheraga, *J. Phys. Chem. B* **108**, 9421 (2004).
- [49] E.-H. Yap, N. L. Fawzi, and T. Head-Gordon, *Proteins* **70**, 626 (2008).
- [50] P. Májek and R. Elber, *Proteins* **76**, 822 (2009).
- [51] D. Alemani, F. Collu, M. Cascella, and M. Dal Peraro, *J. Chem. Theory Comput.* **6**, 315 (2010).
- [52] M. Enciso and A. Rey, *J. Chem. Phys.* **132**, 235102 (2010).
- [53] M. Enciso and A. Rey, *J. Chem. Phys.* **136**, 215103 (2012).
- [54] P. Kar, S. M. Gopal, Y.-M. Cheng, A. Predeus, and M. Feig, *J. Chem. Theory Comput.* **9**, 3769 (2013).
- [55] A. Liwo, S. Oldziej, M. R. Pincus, R. J. Wawak, S. Rackovsky, and H. A. Scheraga, *J. Comput. Chem.* **18**, 849 (1997).
- [56] T. E. Ouldridge, A. A. Louis, and J. P. K. Doye, *Phys. Rev. Lett.* **104**, 178101 (2010).
- [57] M. C. Linak, R. Tourdot, and K. D. Dorfman, *J. Chem. Phys.* **135**, 205102 (2011).
- [58] P. Sulc, F. Romano, T. E. Ouldridge, L. Rovigatti, J. P. K. Doye, and A. A. Louis, *J. Chem. Phys.* **137**, 135101 (2012).
- [59] M. Orsi and J. W. Essex, *PLoS ONE* **6**, e28637 (2011).
- [60] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, *Nucleic Acids Res.* **28**, 235 (2000).
- [61] A. Morriss-Andrews, J. Rottler, and S. S. Plotkin, *J.*

- Chem. Phys. **132**, 035105 (2010).
- [62] E. Spiga, D. Alemani, M. T. Degiacomi, M. Cascella, and M. Dal Peraro, *J. Chem. Theory Comput.* **9**, 3515 (2013).
 - [63] T. K. Haxton, R. V. Mannige, R. N. Zuckermann, and S. Whitelam, submitted.
 - [64] S. M. Gopal, S. Mukherjee, Y.-M. Cheng, and M. Feig, *Proteins* **78**, 1266 (2009).
 - [65] A. Liwo, C. Czaplewski, J. Pillardy, and H. A. Scheraga, *J. Chem. Phys.* **115**, 2323 (2001).
 - [66] Z. Li, Y. Yang, J. Zhan, L. Dai, and Y. Zhou, *Annu. Rev. Biophys.* **42**, 315 (2013).
 - [67] W. Humphrey, A. Dalke, and K. Schulten, *J. Molec. Graphics* **14**, 33 (1996).
 - [68] A. Grishaev, V. Tugurinov, L. E. Kay, J. Trehwella, and A. Bax, *J. Biomol. NMR* **40**, 95 (2008).
 - [69] S. J. Marrink, H. J. Risselada, S. Yefimov, D. P. Tieleman, and A. H. de Vries, *J. Phys. Chem. B* **111**, 7812 (2007).
 - [70] L. Monticelli, S. K. Kandasamy, X. Periole, R. G. Larson, D. P. Tieleman, and S.-J. Marrink, *J. Comput. Chem.* **4**, 819 (2008).
 - [71] Y. Kobashigawa, M. Sakai, M. Naito, M. Yokochi, H. Kumeta, Y. Makino, K. Ogura, S. Tanaka, and F. Inagaki, *Nat. Struct. Mol. Biol.* **14**, 503 (2007).
 - [72] P. Kar, S. M. Gopal, Y.-M. Cheng, A. Panahi, and M. Feig, *J. Chem. Theory Comput.* **10**, 3459 (2014).
 - [73] Y. He, Y. Xiao, A. Liwo, and H. A. Scheraga, *J. Comput. Chem.* **30**, 2127 (2009).
 - [74] R. D. Hills, L. Lu, and G. A. Voth, *PLOS Comput. Biol.* **6**, e1000827 (2010).